SOURCE AND RELIABILITY STATEMENT FOR THE SURVEY OF INCOME AND PROGRAM PARTICIPATION (SIPP) 1985 PUBLIC USE FILES

DATA COLLECTION AND ESTIMATION

Source of Data. The data were collected in the 1985 panel of the Survey of Income and Program Participation (SIPP). The SIPP universe is the noninstitutionalized resident population living in the United States. This population includes persons living in group quarters, such as dormitories, rooming houses, and religious group dwellings. Crew members of merchant vessels, Armed Forces personnel living in military barracks, and institutionalized persons, such as correctional facility inmates and nursing home residents, were not eligible to be in the survey. Also, United States citizens residing abroad were not eligible to be in the survey. Foreign visitors who work or attend school in this country and their families were eligible; all others were not eligible to be in the survey. With the exceptions noted above, persons who were at least 15 years of age at the time of the interview were eligible to be in the survey.

The 1985 panel SIPP sample is located in 230 Primary Sampling Units (PSUs) each consisting of a county or a group of contiguous counties. Within these PSUs, expected clusters of 2 or 4 living quarters (LQs) were systematically selected from lists of addresses prepared for the 1980 decennial census to form the bulk of the sample. To account for LQs built within each of the sample areas after the 1980 census, a sample was drawn of permits issued for construction of residential LQs up until shortly before the beginning of the panel. In jurisdictions that do not issue building permits, small land areas were sampled and the LQs within were listed by field personnel and then subsampled. In addition, sample LQs were selected from supplemental frames that included LQs identified as missed in the 1980 census and group quarters.

Approximately 17,800 living quarters were originally designated for the sample. For Wave 1, interviews were obtained from the occupants of about 13,400 of the 17,800 designated living quarters. Most of the remaining 4,400 living quarters were found to be vacant, demolished, converted to nonresidential use, or otherwise ineligible for the survey. However, approximately 1,000 of the 4,400 living quarters were not interviewed because the occupants refused to be interviewed, could not be found at home, were temporarily absent, or were otherwise unavailable. Thus, occupants of about 93 percent of all eligible living quarters participated in Wave 1 of the survey. For Wave 5, occupants of about 82 percent of all eligible living quarters participated in the survey.

For Waves 2-8, only original sample persons (those in Wave 1 sample households and interviewed in Wave 1 and/or 2) and persons living with them were eligible to be interviewed. With certain restrictions, original sample persons were to be followed even if they moved to a new address. When original sample persons moved without leaving a forwarding address or moved to extremely remote parts of the country and no telephone number was available, additional noninterviews resulted.

Sample households within a given panel are divided into four subsamples of nearly equal size. These subsamples are called rotation groups 1, 2, 3, or 4 and one rotation group is interviewed each month. Each household in the sample was scheduled to be interviewed at 4 month intervals over a period of roughly 2 1/2 years beginning in February 1985. The reference period for the questions is the 4-month period preceding the interview month. In general, one cycle of four interviews covering the entire sample, using the same questionnaire, is called a wave. The exception is Wave 2 which covers three interviews.

The public use files include core and supplemental (topical module) data. Core questions are repeated at each interview over the life of the panel. Topical modules include questions which are asked only in certain waves. The 1985 panel topical modules are given in Table 1.

Table 2 indicates the reference months and interview month for the collection of data from each rotation group for the 1985 panel. For example, Wave 1 rotation group 2 was interviewed in February 1985 and data for the reference months October 1984 through January 1985 were collected.

Table 1. 1985 Panel Topical Modules

Wave	Topical Module
1	None
2	None
3	Assets Liabilities
4	Marital History Fertility History Migration History Household Relationships Support for Non-household Members Work Related Expenses
5	Annual Income Taxes Individual Retirement Accounts Educational Financing and Enrollment
6	Child Care Arrangements Child Support Agreements Support for Non-household Members Job Offers Health Status and Utilization of Health Care Services Long-Term Care Disability Status of Children
7	Assets Liabilities Pension Plan Coverage Lump Sum Distributions from Pension Plans Characteristics of Job from which Retired Characteristics of Home Financing Arrangements
8	Annual Income Taxes Individual Retirement Accounts Educational Financing and Enrollment

Table 2. Reference Months for Each Interview Month - 1985 Panel

Month of Inter view	Wave/ Rota- tion	Reference Period																					
		4th Quarter (1984)			1st Quarter (1985)			2nd Quarter (1985)			3rd Quarter (1985)			4th Quarter (1985)			. . .	2nd Quarter (1987)			3rd Quarter (1987)		
		Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec		Apr	May	Jun	Jul	Aug	Sep
Feb. 85	1/2	X	X	X	X																		
March	1/3		X	X	X	X																	
April	1/4			X	X	X	X																
May	1/1				X	X	X	X															
June	2/2					X	X	X	X														
July	2/3						X	X	X	X													
Aug.	2/4							X	X	X	X												
Sept.	3/1								X	X	X	X											
Oct.	3/2									X	X	X	X										
Nov.	3/3										X	X	X	X									
Dec.	3/4											X	X	X	X								
.														.	.	.							
.																	. . . .						
.																	. . . .						
Aug. 87	8/4																	X	X	X	X		

Assignment of Weights. The estimation procedure used to derive the SIPP person weights involves several stages. These include determining the base weight, adjusting for movers and noninterviews, adjusting to account for the SIPP sample areas not having the same population distribution as the strata from which they were selected and adjusting persons' weights to bring sample estimates into agreement with independent population estimates.

Each person received a base weight equal to the inverse of his/her probability of selection. The SIPP base weight W indicates that each SIPP sample person represents approximately W persons in the SIPP universe. Beginning in Wave 4, base weights were adjusted to account for a February 1986 (Wave 4, rotation 2) sample cut implemented for budgetary reasons. It dropped about 2,000 eligible housing units from the sample. Noninterviews as well as interviews were subject to the cut. In some instances, the base weight was also adjusted to reflect subsampling done in the field. For each subsequent interview, each person received a base weight that accounted for following movers.

A noninterview adjustment factor was applied to the weight of each interviewed person to account for persons in noninterviewed occupied living quarters which were eligible for the sample. (Individual nonresponse within partially interviewed households was treated with imputation. No special adjustment was made for noninterviews in group quarters.) A first stage ratio estimate factor was applied to each interviewed person's weight to account for the SIPP sample areas not having the same population distribution as the strata from which they were selected. In particular, the first stage ratio estimate factors make adjustments by region, race, and by metropolitan and non-metropolitan residence defined as of June 1984.

An additional stage of adjustment to persons' weights was performed to reduce the mean square error of the survey estimates. This was accomplished by bringing the sample estimates into agreement with independent monthly estimates of the civilian (and some military) noninstitutional population of the United States by age, race, Spanish origin, and sex and with special Current Population Survey (CPS) estimates of the prevalence of different types of householders (married, single with relatives or single without relatives by sex and race) and different relationships to householders (spouse or other). The independent estimates were based on statistics from the 1980 Decennial Census of Population; statistics on births, deaths, immigration and emigration; and statistics on the strength of the Armed Forces. Also, husbands and wives were assigned equal weights. As a result of these adjustments, the following types of consistency are attained by race and sex on a monthly basis:

1. The sum of weights of civilian (and some military) noninstitutionalized persons agrees with independent estimates by age-race-Spanish origin-sex groups.

2. The sum of weights of civilian (and some military) noninstitutionalized persons is within a close tolerance of special CPS estimates by householder type and relationship to householder. (The special CPS estimates are similar but not identical to the monthly CPS estimates.)

3. Husbands and wives living together have equal weights. Thus, if a characteristic is necessarily shared by a husband and wife (such as size of family), then the sample estimate of the number of husbands with the characteristic will agree with the corresponding estimate for wives.

Two sources of error were identified in weighting of the 1985 panel. Two first stage factors were incorrect and inconsistent independent controls (independent estimates) were used during the second stage ratio adjustment procedure. The impact of these two error sources on primary SIPP estimates is believed to be minimal.

The first stage factors used for Blacks not in a Metropolitan Statistical Area (MSA) in the Midwest and for non-Blacks not in an MSA in the Midwest were incorrect. If the correct factors were used, it is expected that totals at the national level would be less than 1 percent higher while the impact on the estimated number of Blacks with a given characteristic will be negligible. Totals for non-Blacks at the national level, for the population not in an MSA, and for non-Blacks in the Midwest would exhibit an increase of about 2 percent and totals for non-Blacks not in an MSA in the Midwest would be about 7 percent higher. Since the farm population is heavily concentrated in areas not in an MSA in the Midwest, farm population estimates would be most affected by the errors in the first stage factors. Note that these effects would be observed with estimates based on weights after the first stage adjustment. As a result of second stage weighting adjustments, the effects will be decreased.

Independent control counts (independent estimates) of total population and Hispanics by reference month used during the second stage ratio adjustment portion of the weighting are meant to be consistent. However, the October, November, and December 1985 controls for Hispanics included illegal aliens while those for the total population did not. Total estimates based on these inconsistent controls compared to estimates based on controls without illegal aliens will not be affected. For monthly and quarterly estimates, non-Hispanic totals will be less than 0.3 percent lower, totals for Hispanics and Hispanic males will be about 4 percent higher, and totals for male Hispanics between the ages of 15 and 24 will increase by about 8 percent. For Wave 3 and annual estimates, non-Hispanic totals will be less than 0.1 percent lower, totals for Hispanics and Hispanic males will be about 1 percent higher, and totals for male Hispanics between the ages of 15 and 24 will increase by less than 2 percent. The effects on Wave 4 estimates will be between the Wave 3 and annual and the monthly and quarterly estimate effects.

Use of Weights. Each household and each person within each household on each wave tape has five weights. Four of these weights are reference month specific and therefore can be used only to form reference month estimates. To form an estimate for a particular month, use the **reference month** weight for the month of interest, summing over all persons or households with the characteristic of interest whose reference period includes the month of interest. Multiply the sum by a factor to account for the number of rotations contributing data for the month. This factor equals four divided by the number of rotations contributing data for the month. For example, December 1984 data is only available from rotations 2, 3, and 4 for Wave 1, so a factor of 4/3 must be applied. January 1985 data is available from all four rotations for Wave 1, so a factor of 4/4 = 1 must be applied. Reference month estimates can be averaged to form estimates of monthly averages over some period of time. For example, using the proper weights, one can estimate the monthly average number of households in a specified income range over November and December l984 from Wave 1. The remaining weight is interview month specific. This weight can be used to form estimates that specifically refer to the interview month (e.g., total persons currently looking for work), as well as estimates referring to the time period including the interview month and all previous months (e.g., total persons who have ever served in the military). These tapes contain no weight for characteristics that involve a person's or household's status over two or more months (e.g., number of households with a 50 percent increase in income between November and December l984).

When estimates for months without four rotations worth of data are constructed from a wave file, factors greater than 1 must be applied. However, when core data from consecutive waves are used together, data from all four rotations may be available, in which case the factors are equal to 1.

To estimate monthly averages of a given measure (e.g., total, mean) over a number of consecutive months, sum the monthly estimates and divide by the number of months.

Producing Estimates for Census Regions and States. The total estimate for a region is the sum of the state estimates in that region.

Estimates from this sample for individual states are subject to very high variance and are not recommended. The state codes on the file are primarily of use for linking respondent characteristics with appropriate contextual variables (e.g., state-specific welfare criteria) and for tabulating data by user-defined groupings of states.

Producing Estimates for the Metropolitan Population. For Washington, DC and 11 states, metropolitan or non-metropolitan residence is identified (variable H*-METRO, characters 94, 382, 670, and 958). In 34 additional states, where the non-metropolitan population in the sample was small enough to present a disclosure risk, a fraction of the metropolitan sample was recoded so as to be indistinguishable from non-metropolitan cases (H*-METRO=2). In these states, therefore, the cases coded as metropolitan (H*-METRO=1) represent only a subsample of that population.

In producing state estimates for a metropolitan characteristic, multiply the individual, family, or household weights by the metropolitan inflation factor for that state, presented in Table 6. (This inflation factor compensates for the subsampling of the metropolitan population and is 1.0 for the states with complete identification of the metropolitan population.) The same procedure applies when creating estimates for particular identified MSA's or CMSA's - apply the factor appropriate to the state. For multi-state MSA's, use the factor appropriate to each state part. For example, to tabulate data for the Washington, DC-MD-VA MSA, apply the Virginia factor of 1.0521 to weights for residents of the Virginia part of the MSA; Maryland and DC residents require no modification to the weights (i.e., their factors equal 1.0).

In producing regional or national estimates of the metropolitan population, it is also necessary to compensate for the fact that no metropolitan subsample is identified within two states (Mississippi and West Virginia) and one state-group (North Dakota - South Dakota - Iowa). Thus, factors in the right-hand column of Table 6 should be used for regional and national estimates. The results of regional and national tabulations of the metropolitan population will be biased slightly. However, less than one-half of one percent of the metropolitan population is not represented.

Producing Estimates for the Non-Metropolitan Population. State, regional, and national estimates of the non-metropolitan population cannot be computed directly, except for Washington, DC and the 11 states where the factor for state tabulations in Table 6 is 1.0. In all other states, the cases identified as not in the metropolitan subsample (METRO=2) are a mixture of non-metropolitan and metropolitan households. Only an indirect method of estimation is available: first compute an estimate for the total population, then subtract the estimate for the metropolitan population. The results of these tabulations will be slightly biased.

RELIABILITY OF THE ESTIMATES

SIPP estimates obtained from the public use files are based on a sample; they may differ somewhat from the figures that would be obtained if a complete census had been taken using the same questionnaire, instructions, and enumerators. There are two types of errors possible in an estimate based on a sample survey: nonsampling and sampling. The magnitude of SIPP sampling error can be estimated, but this is not true of nonsampling error. Found below are descriptions of sources of SIPP non-sampling error, followed by a discussion of sampling error, its estimation, and its use in data analysis.

Nonsampling Variability. Nonsampling errors can be attributed to many sources, e.g., inability to obtain information about all cases in the sample, definitional difficulties, differences in the interpretation of questions, inability or unwillingness on the part of the respondents to provide correct information, inability to recall information, errors made in collection such as in recording or coding the data, errors made in processing the data, errors made in estimating values for missing data, biases resulting from the differing recall periods caused by the rotation pattern used and failure to represent all units within the universe (undercoverage). Quality control and edit procedures were used to reduce errors made by respondents, coders and interviewers.

Undercoverage in SIPP results from missed living quarters and missed persons within sample households. It is known that undercoverage varies with age, race, and sex. Generally, undercoverage is larger for males than for females and larger for Blacks than for nonblacks. Ratio estimation to independent age-race-Spanish origin-sex population controls partially corrects for the bias due to survey undercoverage. However, biases exist in the estimates to the extent that persons in missed households or missed persons in interviewed households have different characteristics than the interviewed persons in the same age-race-Spanish origin-sex group. Further, the independent population controls used have not been adjusted for undercoverage in the decennial census.

The following table summarizes information on household nonresponse for the interview months for Wave 1.

Sample Size, by Month and Interview Status

	Household Units Eligible			
Month	Total	Inter-viewed	Not Inter-viewed	Non-Response Rate (%)
Feb 1985	3,500	3,300	300	7
Mar 1985	3,600	3,400	200	6
Apr 1985	3,600	3,400	200	6
May 1985	3,600	3,300	300	7

Due to rounding of all numbers at 100, there are some inconsistencies. The non-response rate was calculated using unrounded numbers.

Additional noninterviews and the sample cut implemented in February 1986, resulted in the interviewed sample size decreasing to about 10,800 for Wave 5. Sample loss at Wave 1 was about 7 percent and increased to roughly 19 percent at the end of Wave 5. Further non-interviews increased the sample loss about 1 percent for each of the remaining waves.

Some respondents do not respond to some of the questions. Therefore, the overall nonresponse rate for some items such as income and other money related items is higher than the nonresponse rates in the above table. The Bureau has used complex techniques to handle nonresponse, but the success of these techniques in avoiding the bias resulting from overall nonresponse is unknown.

Comparability with other statistics. Caution should be exercised when comparing data from these files with data from other SIPP products or with data from other surveys. The comparability problems are caused by the seasonal patterns for many characteristics and by different nonsampling errors.

Sampling variability. Standard errors indicate the magnitude of the sampling error. They also partially measure the effect of some nonsampling errors in response and enumeration, but do not measure any systematic biases in the data. The standard errors for the most part measure the variations that occurred by chance because a sample rather than the entire population was surveyed.

Confidence Intervals. The sample estimate and its standard error enable one to construct confidence intervals, ranges that would include the average result of all possible samples with a known probability. For example, if all possible samples were selected, each of these being surveyed under essentially the same conditions and using the same sample design, and if an estimate and its standard error were calculated from each sample, then approximately 90 percent of the intervals from 1.6 standard errors below the estimate to 1.6 standard errors above the estimate would include the average result of all possible samples.

The average estimate derived from all possible samples is or is not contained in any particular computed interval. However, for a particular sample, one can say with a specified confidence that the average estimate derived from all possible samples is included in the confidence interval.

Hypothesis Testing. Standard errors may also be used for hypothesis testing, a procedure for distinguishing between population parameters using sample estimates. The most common types of hypotheses tested are

1) the population parameters are identical versus 2) they are different. Tests may be performed at various levels of significance, where a level of significance is the probability of concluding that the parameters are different when, in fact, they are identical.

To perform the most common test, let x and y be sample estimates of two parameters of interest. A subsequent section explains how to derive a standard error on the difference x-y. If the estimated absolute difference between parameters is greater than 1.6 times the standard error of the difference, then the observed difference is significant at the 10 percent level. In this event, it is commonly accepted practice to say that the parameters are different. Of course, sometimes this conclusion will be wrong. When the parameters are, in fact, the same, there is a 10 percent chance of concluding that they are different. We recommend that users report only those differences that are significant at the 10 percent level or better.

Note when using small estimates. Because of the large standard errors involved, there is little chance that estimates will reveal useful information when computed on a base smaller than 200,000. Nonsampling error in one or more of the small number of cases providing the estimate can cause large relative error in that particular estimate. Also care must be taken in the interpretation of small differences. For instance, in case of a borderline difference, even a small amount of nonsampling error can lead to a wrong decision about the hypotheses, thus distorting a seemingly valid hypothesis test.

Standard Error Parameters and Tables and Their Use. To derive standard errors that would be applicable to a wide variety of statistics and could be prepared at a moderate cost, a number of approximations were required. Most of the SIPP statistics have greater variance than those obtained through a simple random sample because clusters of living quarters are sampled for the SIPP. Two parameters (denoted "a" and "b") were developed to quantify these variances. These "a" and "b" parameters are used in estimating standard errors of survey estimates. The "a" and "b" parameters vary by type of estimate and by subgroup to which the estimate applies. Table 4 provides base "a" and "b" parameters for various subgroups and types of estimates. The factors provided in Table 5 when multiplied by the base parameters for a given subgroup and type of estimate give the "a" and "b" parameters for that subgroup and estimate type for the specified reference period. For example, the base "a" and "b" parameters for total income of households are -0.0001062 and 9407, respectively. For Wave 1, the factor for October l984 is 4 since only 1 rotation of data is available. So, the "a" and "b" parameters for total household income in October l984 based on Wave 1 are -0.0004248 and 37,628, respectively. Also for Wave 1, the factor for the first quarter of 1985 is 1.2222 since 9 rotation months of data are available (rotations 1 and 4 provide 3 rotation months each, while rotations 2 and 3 provide 1 and 2 rotation months, respectively). So, the "a" and "b" parameters for total household income in the first quarter of 1985 are -0.0001298 and 11,497, respectively for Wave 1.

The "a" and "b" parameters may be used to directly calculate the standard error for estimated numbers and percentages. Because the actual variance behavior was not identical for all statistics within a group, the standard errors computed from these parameters provide an indication of the order of magnitude of the standard error for any specific statistic. Methods for using these parameters for direct computation of standard errors are given in the following sections.

Procedures for calculating standard errors for the types of estimates most commonly used are described below. Note specifically that these procedures apply only to reference month estimates or averages of reference month estimates. Refer to the section "Use of Weights" for a detailed discussion of construction of estimates. Stratum codes and half sample codes are included on the tapes to enable the user to compute the variances directly by methods such as balanced repeated replications (BRR). William G. Cochran provides a list of references discussing the application of this technique.[1]

Standard errors of estimated numbers. The approximate standard error of an estimated number can be obtained by using formula (1).

---

1. Cochran, William G. (1977), *Sampling Techniques*, 3rd Edition, New York: John Wiley and Sons, p.321.

$$s_x = \sqrt{ax^2 + bx} \qquad (1)$$

Here x is the size of the estimate and "a" and "b" are the parameters associated with the particular type of characteristic for the appropriate reference period.

Illustration. Suppose that the SIPP estimates from Wave 1 show an estimated 31,555,000 persons in non-farm households with a mean monthly household cash income of $4,000 or over during January 1985 for which four rotations of data are available. Then the appropriate base "a" and "b" parameters and factor to use in calculating a standard error for the estimate are obtained from tables 4 and 5. They are a = -0.0000446 and b = 7612 with a factor of 1.0.

Using formula (1), the approximate standard error is

$$\sqrt{(-0.0000446)\ (31{,}555{,}000)^2 + (7612)\ (31{,}555{,}000)} \approx 442{,}479$$

The 90-percent confidence interval as shown by the data is from 30,847,034 to 32,262,966.

Standard errors of estimated percentages. This section refers to percentages of a group of persons, families, or households possessing a particular attribute (e.g., the percentage of households receiving food stamps).

The reliability of an estimated percentage, computed using sample data for both numerator and denominator, depends upon both the size of the percentage and the size of the total upon which the percentage is based. Estimated percentages are relatively more reliable than the corresponding estimates of the numerators of the percentages, particularly if the percentages are 50 percent or more, e.g., the percent of people employed. When the numerator and denominator of the percentage have different parameters, use the parameters for the numerator. The approximate standard error, $s_{x,p}$, of the estimated percentage p can be obtained by the formula

$$s_{x,p} = \sqrt{\frac{b}{x}\ (p[100-p])} \qquad (2)$$

Here x is the size of the subclass of households or persons in households which is the base of the percentage, p is the percentage ($0<p<100$), and b is the "b" parameter for the numerator.

Illustration. Continuing the example from above, suppose Wave 1 data shows that of the 31,555,000 persons in non-farm households with a mean monthly household cash income of $4,000 or over, 91.9 percent were White. Using formula (2) and the appropriate base "b" parameter and factor from tables 4 and 5, the approximate standard error is

$$\sqrt{\frac{(7{,}612)}{(31{,}555{,}000)}\ (91.9)\ (100-91.9)} \approx 0.4 \text{ percent}$$

Consequently, the 90 percent confidence interval as shown by these data is from 91.3 to 92.5 percent.

Standard error of a mean. A mean is defined here to be the average quantity of some item (other than persons, families, or households) per person, family, or household. (For the mean of these other items, compute the standard error using formula (9).) For example, the mean could be the average monthly household income of females age 25 to 34. The standard error of such a mean can be approximated by formula (3) below. Because of the approximations used in developing formula (3), an estimate of the standard error of the mean obtained from that formula will generally underestimate the true standard error. The formula used to estimate the standard error of a mean $\bar{x}$ is

$$\sqrt{s_{\bar{x}} = \frac{b}{y}\ s^2} \qquad (3)$$

where y is the size of the base, s^2 is the estimated population variance of the item and b is the parameter associated with the particular type of item.

The estimated population variance, s^2, is given by:

$$s^2 = \sum_{i=1}^{c} p_i x_i^2 - \bar{x}^2 \qquad (4)$$

$$\text{where: } \bar{x} = \sum_{i=1}^{c} p_i x_i . \qquad (5)$$

It is assumed that each person or other unit was placed in one of c groups; p_i is the estimated proportion of group i; $x_i = (Z_{i-1} + Z_i)/2$ where Z_{i-1} and Z_i are the lower and upper interval boundaries, respectively, for group i. x_i is assumed to be the most representative value for the characteristic of interest in group i. If group c is open-ended, i.e., no upper interval boundary exists, then an approximate value for x_c is

$$x_c = \frac{3}{2} Z_{c-1} \qquad (6)$$

Illustration. Suppose that based on Wave 1 data, the distribution of monthly income for persons age 25 to 34 during January l985 is given in the following table.

Table 3. Distribution of Monthly Income Among Persons 25 To 34 Years Old.

	Total	Under $300	$300 to $599	$600 to $899	$900 to $1,199	$1,200 to $1,499	$1,500 to $1,999	$2,000 to $2,499	$2,500 to $2,999	$3,000 to $3,499	$3,500 to $3,999	$4,000 to $4,999	$5,000 to $5,999	$6,000 and over
Thousands in interval	39,851	1371	1651	2259	2734	3452	6278	5799	4730	3723	2519	2619	1223	1493
Percent with at least as much as lower bound of interval	--	100.0	96.6	92.4	86.7	79.9	71.2	55.5	40.9	29.1	19.7	13.4	6.8	3.7

Using formula (4) and the mean monthly cash income of $2,530 the approximate population variance, s^2, is

$$s^2 = \frac{1{,}371}{39{,}851} (150)^2 + \frac{1{,}651}{39{,}851} (450)^2 + \ldots..$$

$$+ \frac{1{,}493}{39{,}851} (9{,}000)^2 - (2{,}530)^2 = 3{,}159{,}887.$$

Using formula (3), the appropriate base "b" parameter and factor, the estimated standard error of a mean $\bar{x}$ is

$$s_{\bar{x}} = \sqrt{\frac{7,612}{39,851,000}\quad(3,159,887)} = \$25$$

Standard error of a median. The median quantity of some item such as income for a given group of persons, families, or households is that quantity such that at least half the group have as much or more and at least half the group have as much or less. The sampling variability of an estimated median depends upon the form of the distribution of the item as well as the size of the group. An approximate method for measuring the reliability of an estimated median is to determine a confidence interval about it. (See the section on sampling variability for a general discussion of confidence intervals.) The following procedure may be used to estimate the 68-percent confidence limits and hence the standard error of a median based on sample data.

1. Determine, using formula (2), the standard error of an estimate of 50 percent of the group;

2. Add to and subtract from 50 percent the standard error determined in step (1);

3. Using the distribution of the item within the group, calculate the quantity of the item such that the percent of the group owning more is equal to the smaller percentage found in step (2). This quantity will be the upper limit for the 68-percent confidence interval. In a similar fashion, calculate the quantity of the item such that the percent of the group owning more is equal to the larger percentage found in step (2). This quantity will be the lower limit for the 68-percent confidence interval;

4. Divide the difference between the two quantities determined in step (3) by two to obtain the standard error of the median.

To perform step (3), it will be necessary to interpolate. Different methods of interpolation may be used. The most common are simple linear interpolation and Pareto interpolation. The appropriateness of the method depends on the form of the distribution around the median. If density is declining in the area, then we recommend Pareto interpolation. If density is fairly constant in the area, then we recommend linear interpolation. Note, however, that Pareto interpolation can never be used if the interval contains zero or negative measures of the item of interest. Interpolation is used as follows. The quantity of the item such that "p" percent own more is

$$X_{pN} = A_1 \exp\left[Ln\left(\frac{pN}{N_1}\right) \quad Ln\left(\frac{A_2}{A_1}\right) \Big/ Ln\left(\frac{N_2}{N_1}\right)\right] \tag{7}$$

if Pareto interpolation is indicated and

$$X_{pN} = \frac{N_1 - pN}{N_1 - N_2} \quad (A_2 - A_1) \quad + \quad A_1 \tag{8}$$

if linear interpolation is indicated,

where

N	is size of the group,
A_1 and A_2	are the lower and upper bounds, respectively, of the interval in which X_{pN} falls,
N_1 and N_2	are the estimated number of group members owning more than A_1 and A_2, respectively,

exp	refers to the exponential function, and
Ln	refers to the natural logarithm function.

It should be noted that a mathematically equivalent result is obtained by using common logarithms (base 10) and antilogarithms.

Illustration. To illustrate the calculations for the sampling error on a median, we return to the same example used to illustrate the standard error of a mean. The median monthly income for this group is $2,158. The size of the group is 39,851,000.

1. Using formula (2), the standard error of 50 percent on a base of 39,851,000 is about .7 percentage points.

2. Following step (2), the two percentages of interest are 49.3 and 50.7.

3. By examining Table 3, we see that the percentage 49.3 falls in the income interval from $2,000 to $2,499. (Since 55.5 percent receive more than $2,000 per month, but only 40.9 percent receive more than $2,500 per month, the dollar value corresponding to 49.3 percent must be between $2,000 and $2,500.) Thus A_1 = $2,000, A_2 = $2,500, N_1 = 22,106,000, and N_2 = 16,307,000.

In this case, we decided to use Pareto interpolation. Therefore, the upper bound of a 68-percent confidence interval for the median is

$$\$2{,}000 \exp \left[\text{Ln} \left(\frac{(.493)(39{,}851{,}000)}{22{,}106{,}000} \right) \text{Ln}\left(\frac{2{,}500}{2{,}000}\right) \Big/ \text{Ln} \left(\frac{16{,}307{,}000}{22{,}106{,}000}\right) \right] = \$2{,}181$$

Also by examining Table 3, we see that 50.7 falls in the same income interval. Thus, A_1, A_2, N_1, and N_2 are the same. We also decided to use Pareto interpolation for this case. So the lower bound of a 68-percent confidence interval for the median is

$$\$2{,}000 \exp \left[\text{Ln} \left(\frac{(.507)(39{,}851{,}000)}{22{,}106{,}000} \right) \text{Ln}\left(\frac{2{,}500}{2{,}000}\right) \Big/ \text{Ln} \left(\frac{16{,}307{,}000}{22{,}106{,}000}\right) \right] = \$2{,}136$$

Thus, the 68-percent confidence interval on the estimated median is from $2,136 to $2,181. An approximate standard error is

$$\frac{\$2{,}181 - \$2{,}136}{2} = \$23.$$

Standard errors of ratios. The standard error for the average quantity of persons, families, or households per family or household or for a ratio of means or medians is approximated by formula (9):

$$s_{\left(\frac{x}{y}\right)} = \sqrt{\left(\frac{x}{y}\right)^2 \left[\left(\frac{s_y}{y}\right)^2 + \left(\frac{s_x}{x}\right)^2\right]} \qquad (9)$$

where x and y are the numerator and denominator for the average or the means or medians which form the ratio, and s_x and s_y are their associated standard errors. Formula (9) assumes that x and y are not correlated. If the correlation is actually positive (negative), then this procedure will provide an overestimate (underestimate) of the standard error for the ratio.

Standard error of a difference. The standard error of a difference between two sample estimates is approximately equal to

$$s_{(x-y)} = \sqrt{s_x^2 + s_y^2} \qquad (10)$$

where s_x and s_y are the standard errors of the estimates x and y. The estimates can be numbers, percents, ratios, etc. The above formula assumes that the sample correlation coefficient, r, between the two estimates is zero. If r is really positive (negative), then this assumption will lead to overestimates (underestimates) of the true standard error.

Illustration. Suppose SIPP estimates based on Wave 1 data show that during the first quarter of 1985 the number of persons age 25-34 years in non-farm households with mean monthly cash income of $4,000 to $4,999 was 2,619,000, while the number with mean monthly cash income of $5,000 to $5,999 was 1,223,000. The standard errors of these numbers would be 155,000 and 106,000, respectively.

Suppose that it is desired to test at the 10 percent significance level whether the number of persons age 25-34 in non-farm households was different for persons with a mean monthly cash income of $4,000 to $4,999 than for persons with mean monthly cash income of $5,000 to $5,999 during the first quarter of 1985. Assuming that these two estimates are not correlated, the standard error of the estimated difference of 1,396,000 is

$$\sqrt{(155{,}000)^2 + (106{,}000)^2} \approx 188{,}000.$$

Since the difference is greater than 1.6 times the standard error of the difference it is concluded that there is a significant difference between the two income categories at the 10 percent significance level.

Combined Panel Estimates. Both the 1984 and 1985 panels provide data for October 1984 - July 1986. Thus, estimates made within this time period may be obtained by combining the panels. However, since the Wave 1 questionnaire differs from the subsequent waves' questionnaires and since there were some procedural changes between the 1984 and 1985 panels, we recommend that estimates from Wave 1 of the 1985 panel not be combined with 1984 panel estimates. Additionally, even for later waves, care should be taken when combining data from the two panels since questionnaires for the two panels differ somewhat.

Starting with Wave 2 of the 1985 panel, corresponding data from the 1984 and 1985 panels can be combined to create joint estimates of level by using the formula:

$$\hat{x} = f\,\hat{y} + (1-f)\,\hat{z} \qquad (11)$$

where:

$\hat{x}$ = joint estimate of level;

$\hat{y}$ = estimate of level from the 1984 panel ;

$\hat{z}$ = estimate of level from the 1985 panel ;

f = 1984 panel weighting factor. The following values should be used when combining data from rotations for the given waves.

Waves to be combined

1985 panel	1984 panel	f
2*	6	.546
3	7	.543
4*	8	.566
5*	9	.566

*For these waves, only three rotations overlap the corresponding wave of the 1984 panel.

The approximate standard error of the combined estimate $(\hat{x})$ is:

$$S_{\hat{x}} = \sqrt{f^2 \ (S_{\hat{y}})^2 + (1 - f)^2 \ (S_{\hat{z}})^2}$$

where $S_{\hat{x}}$, $S_{\hat{y}}$, and $S_{\hat{z}}$ are the standard errors for the estimates of level for the 1984 and 1985 panels combined, the 1984 panel and the 1985 panel, respectively.

Joint estimates of the more complex statistics (proportions, means, medians, etc.) for a particular characteristic should be calculated from a joint distribution of the characteristic which can be obtained as follows. Generate separate cumulative distributions for the characteristic based on 1984 and 1985 panel data using the same intervals for both distributions. Create a joint distribution by averaging the estimates of level within each interval using formula (11). The complex statistics can then be calculated from the resulting joint distribution.

Table 4. SIPP INDIRECT GENERALIZED VARIANCE PARAMETERS FOR THE 1985 PANEL PUBLIC USE FILE[1]

CHARACTERISTICS	a	b
PERSONS		
Total or White		
16+ Program Participation and Benefits, Poverty (3)		
Both Sexes	-0.0001311	22,327
Male	-0.0002758	22,327
Female	-0.0002497	22,327
16+ Income and Labor Force (5)		
Both Sexes	-0.0000446	7,612
Male	-0.0000941	7,612
Female	-0.0000851	7,612
16+ Pension Plan[2] (4)		
Both Sexes	-0.0000817	13,940
Male	-0.0001723	13,940
Female	-0.0001558	13,940
All Others[2] (6)		
Both Sexes	-0.0001201	27,683
Male	-0.0002483	27,683
Female	-0.0002325	27.683
Black		
Poverty (1)		
Both Sexes	-0.0006903	19,045
Male	-0.0014833	19,045
Female	-0.0012910	19,045
All Others (2)		
Both Sexes	-0.0003712	10,241
Male	-0.0007976	10,241
Female	-0.0006942	10,241
HOUSEHOLDS		
Total or White	-0.0001062	9,407
Black	-0.0006480	6,500

1. Multiply these parameters by 1.35 for estimates which include data from reference month November 1985 and later, except for 1985 calendar year estimates. For calendar year 1985 estimates, use the parameters as given.

   For cross-tabulations, use the parameters of the characteristic with the smaller number within the parentheses.

2. Use the "16+ Pension Plan" parameters for pension plan tabulations of persons 16+ in the labor force. Use the "All Others" parameters for retirement tabulations, 0+ program participation, 0+ benefits, 0+ income, and 0+ labor force tabulations, in addition to any other types of tabulations not specifically covered by another characteristic in this table.

Table 5. Factors to be Applied to Base Parameters to Obtain Parameters for Various Reference Periods

# of available rotation months[1]	factor
Monthly estimate	
1	4.0000
2	2.0000
3	1.3333
4	1.0000
Quarterly estimate	
6	1.8519
8	1.4074
9	1.2222
10	1.0494
11	1.0370
12	1.0000

1. The number of available rotation months for a given estimate is the sum of the number of rotations available for each month of the estimate.

t

Table 6. Metropolitan Subsample Factors to be Applied to Compute National and Subnational Estimates

		Factors for use in State or CMSA (MSA) Tabulations	Factors for use in Regional or National Tabulations
Northeast:	Connecticut	1.0387	1.0387
	Maine	1.2219	1.2219
	Massachusetts	1.0000	1.0000
	New Hampshire	1.2234	1.2234
	New Jersey	1.0000	1.0000
	New York	1.0000	1.0000
	Pennsylvania	1.0096	1.0096
	Rhode Island	1.2506	1.2506
	Vermont	1.2219	1.2219
Midwest:	Illinois	1.0000	1.0110
	Indiana	1.0336	1.0450
	Iowa	--	--
	Kansas	1.2994	1.3137
	Michigan	1.0328	1.0442
	Minnesota	1.0366	1.0480
	Missouri	1.0756	1.0874
	Nebraska	1.6173	1.6351
	North Dakota	--	--
	Ohio	1.0233	1.0346
	South Dakota	--	--
	Wisconsin	1.0188	1.0300
South:	Alabama	1.1574	1.1595
	Arkansas	1.6150	1.6179
	Delaware	1.5593	1.5621
	D.C.	1.0000	1.0018
	Florida	1.0140	1.0158
	Georgia	1.0142	1.0160
	Kentucky	1.2120	1.2142
	Louisiana	1.0734	1.0753
	Maryland	1.0000	1.0018
	Mississippi	--	--
	North Carolina	1.0000	1.0018
	Oklahoma	1.0793	1.0812
	South Carolina	1.0185	1.0203
	Tennessee	1.0517	1.0536
	Texas	1.0113	1.0131
	Virginia	1.0521	1.0540
	West Virginia	--	--
West:	Alaska	1.4339	1.4339
	Arizona	1.0117	1.0117
	California	1.0000	1.0000
	Colorado	1.1306	1.1306
	Hawaii	1.0000	1.0000
	Idaho	1.4339	1.4339
	Montana	1.4339	1.4339
	Nevada	1.0000	1.0000
	New Mexico	1.0000	1.0000
	Oregon	1.1317	1.1317
	Utah	1.0000	1.0000
	Washington	1.0456	1.0456
	Wyoming	1.4339	1.4339

-- indicates no metropolitan subsample is identified for the state